Power Week 2025

#pw2025

18 - 19 - 20 novembre 2025
IBM Innovation Studio Paris

S32 - IA sur IBM Power

18 novembre 16:00 – 17:00

Thibaud BESSON
IBM France
thibaud.besson@fr.ibm.com

Power Week

18 -19 - 20 novembre 2025

FRANCE

Agenda

- Contexte technologique de l’IA
  - Historique et situation actuelle
  - Défis technologiques de l’IA d’entreprise
- Réponse d’IBM
  - Travaux de recherche
  - Carte Spyre
  - Solution Spyre

# En 1993...
# 2D sur CPU
# ou 3D professionnelle

1996 – L’accélération 3D sur GPU

Quake

Tomb Raider

1996 – Accélération 3D

Les Shaders :

- Code dans le GPU qui calcule les géométries et couleurs, en temps réel.
- **Vecteurs** pour les positions et les couleurs,
- **Matrices** pour projeter les positions en 3D dans l’espace 2D de l’écran.

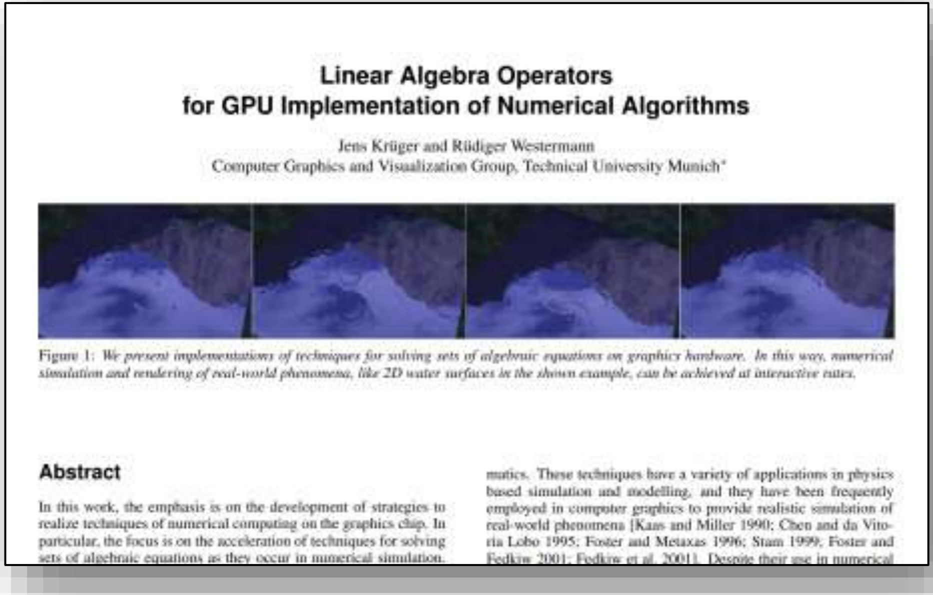
Cartes leader du marché :

- 3dfx Voodoo Graphics
- ATI 3D
- Nvidia NV1

3Dfx Voodoo Graphics Voodoo 1	
Fréquence	50 MHz
Bus	64 bits
RAM dédiée	4 Mo
Résolution	800×600
Communication	PCI
API supportée	Direct3D Shader DirectX3

2003 – GPU en calcul scientifique

- Les **vecteurs** et les **matrices** sont les objets mathématiques de base utilisés en Algèbre linéaire.
- L’**algèbre linéaire** étudie les transformations linéaires entre ces objets.
- Il est fondamental dans de nombreux domaines de la physique et l’ingénierie :
  - ✓ Mécanique quantique : vecteurs d’état dans les espaces de Hilbert, observables en tant qu’opérateurs Hermitiens, solutions de l’équation de Schrödinger.
  - ✓ Mécanique Classique : oscillateurs, mouvements linéaires et rotationnels.
  - ✓ Électromagnétisme : équations de Maxwell.
  - ✓ Traitement du signal : transformées de Fourier, filtres, antennes.
  - ✓ Science des matériaux : contraintes, deformations, analyse structurelle.
  - ✓ Mécanique des Fluides et transfert de chaleur : équation de Navier-Stokes et équation de diffusion thermique
- **Intelligence artificielle** :
  - ✓ les images sont représentées par des tenseurs,
  - ✓ les tokens d’un texte sont des vecteurs dans une base de donnée vectorielle,
  - ✓ les réseaux de neurones du deep learning sont activés selon la matrice de chaque couche, stimulés par un vecteur d’entrée
  - ✓ Le mécanisme d’attention des LLM est codé sous forme de matrices



Linear Algebra Operators for GPU Implementation of Numerical Algorithms

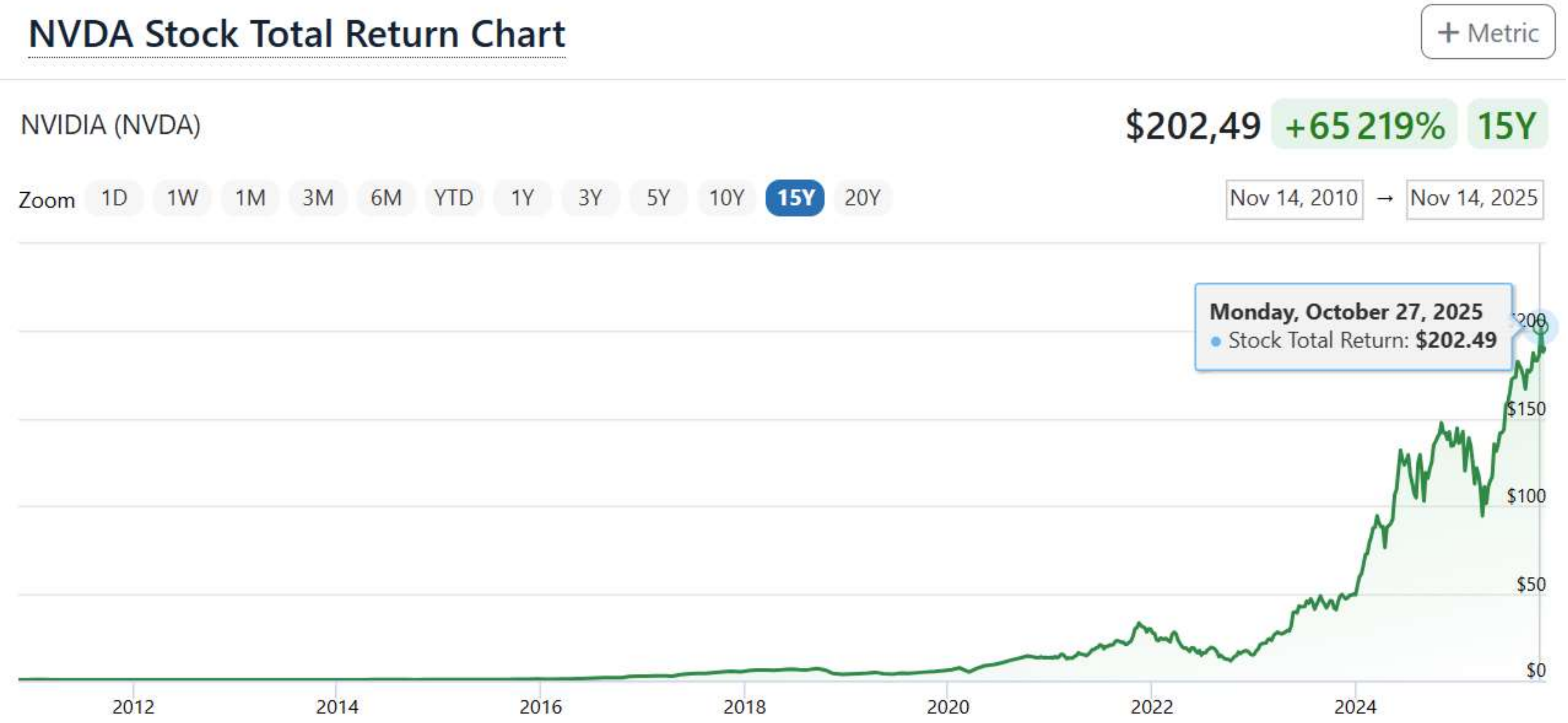
Jens Krüger and Rüdiger Westermann
Computer Graphics and Visualization Group, Technical University Munich*

Figure 1: *We present implementations of techniques for solving sets of algebraic equations on graphics hardware. In this way, numerical simulation and rendering of real-world phenomena, like 2D water surfaces in the shown example, can be achieved at interactive rates.*

Abstract

In this work, the emphasis is on the development of strategies to realize techniques of numerical computing on the graphics chip. In particular, the focus is on the acceleration of techniques for solving sets of algebraic equations as they occur in numerical simulation.

matics. These techniques have a variety of applications in physics based simulation and modelling, and they have been frequently employed in computer graphics to provide realistic simulation of real-world phenomena [Kass and Miller 1990; Chen and da Vitoria Lobo 1995; Foster and Metaxas 1996; Stam 1999; Foster and Fedkiw 2001; Fedkiw et al. 2001]. Despite their use in numerical



NVDA Stock Total Return Chart

NVIDIA (NVDA) $202,49 +65 219% 15Y

Zoom 15Y

Nov 14, 2010 → Nov 14, 2025

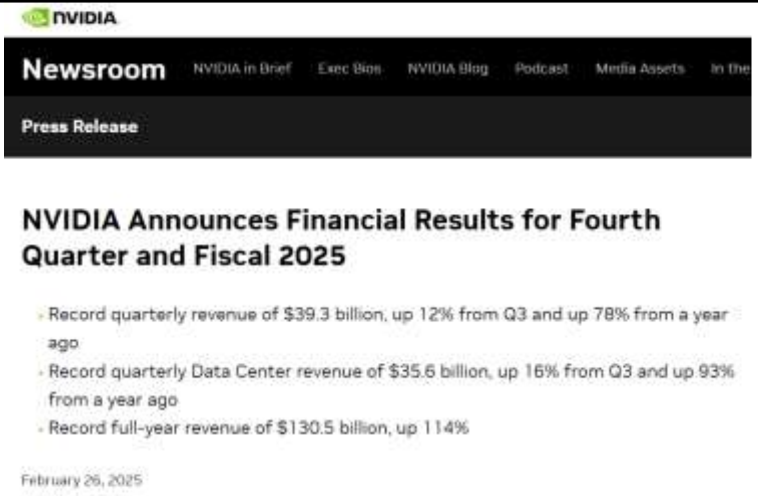
Monday, October 27, 2025
● Stock Total Return: **$202.49**

NVIDIA écrase le marché du GPU

- **Chiffre d'affaires record en 2025** : 130,5 milliards $, en **hausse de 114 %** par rapport à 2024
- Segment Data Center (IA) : 116,2 milliards $, soit **89 % des revenus**
- Deuxième trimestre 2025 : NVIDIA a encore renforcé son avance avec **94 % de parts de marché**, AMD tombant à 6 % et Intel restant sous 1 %
- En volume, 10,9 millions de cartes NVIDIA vendues.

Conséquences du quasi-monopole

- Hausse des **prix** sur les modèles haut de gamme
- Ruptures de **stock** fréquentes
- Risque de ralentissement de **l'innovation** à long terme



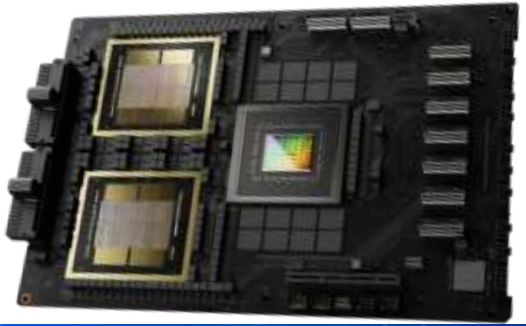
NVIDIA Announces Financial Results for Fourth Quarter and Fiscal 2025

- Record quarterly revenue of $39.3 billion, up 12% from Q3 and up 78% from a year ago
- Record quarterly Data Center revenue of $35.6 billion, up 16% from Q3 and up 93% from a year ago
- Record full-year revenue of $130.5 billion, up 114%

February 26, 2025

Historique de la gamme NVIDIA

GPU Model	CUDA Cores	Tensor Cores	Power Usage (W)	First Availability	Memory Size	Memory Bandwidth	Architecture	Approx Retail Price (USD)	FP32 (TFLOPS)	FP16 (TFLOPS)	Tensor FP16 (TFLOPS)	INT8 (TOPS)
NVIDIA V100	5120	640	250	Jun 2017	16–32 GB HBM2	900 GB/s	Volta	$8,000–$10,000	14.1	28.3	113	56
NVIDIA T4	2560	320	**70**	**Sep 2018**	16 GB GDDR6	320 GB/s	Turing	$600–$900	8.1	16.3	**65**	130
NVIDIA A100	6912	432	400	Jun 2020	80 GB HBM2e	2039 GB/s	Ampere	$9,500–$14,000	19.5	39	312	624
NVIDIA RTX A6000	10752	336	300	Oct 2020	48 GB GDDR6	768 GB/s	Ampere	$4,749–$5,299	38.7	77	309	619
NVIDIA A10	9216	288	150	Apr 2021	24 GB GDDR6	600 GB/s	Ampere	$2,800–$3,300	31.2	62.5	125	250
NVIDIA H100	14592	456	**700**	**Oct 2022**	80 GB HBM3	3000 GB/s	Hopper	$25,000–**$30,000**	51	102	1979	3958
NVIDIA L40S	18176	568	**350**	**Oct 2022**	**48 GB** GDDR6	864 GB/s	Ada Lovelace	$7,500–**$8,750**	91.6	183	**733**	1466
NVIDIA L4	7,424	240	**72**	**Mar 2023**	**24 GB** DDR6	300 GB/s	Ada Lovelace	$2,000–**$2,500**	30.3	30.3	**121***	242*
NVIDIA B200	18944	592	**1000**	**Jan 2024**	**192 GB** HBM3e	8000 GB/s	Blackwell	$45,000–**$50,000**	75	150	**2250**	4500



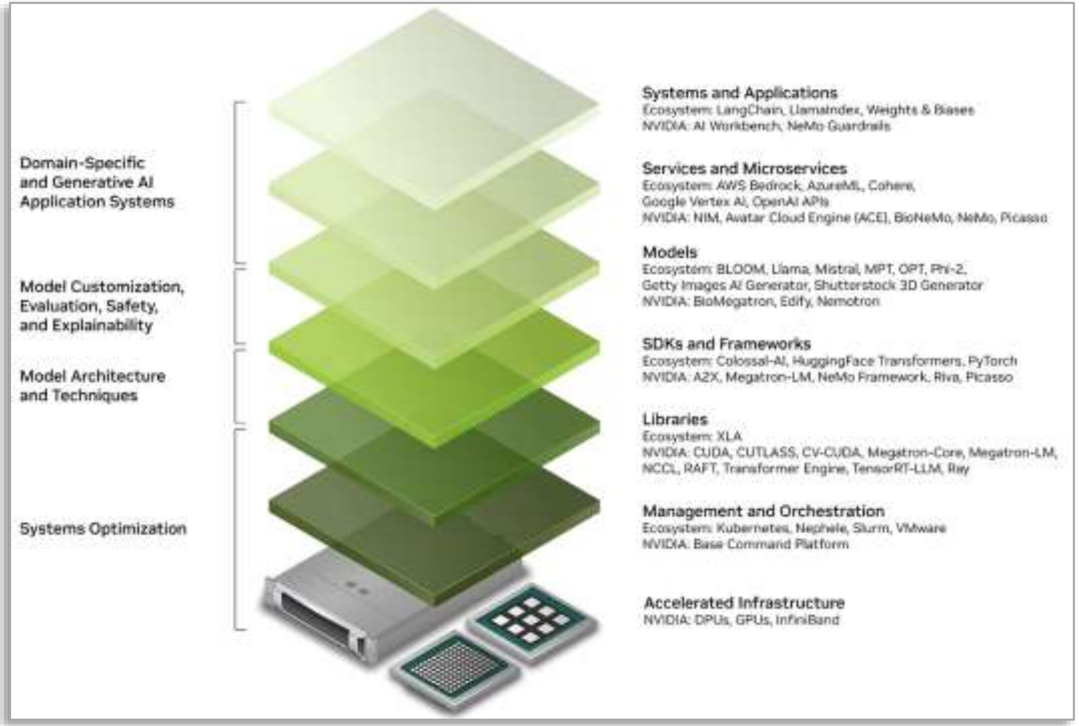
Google TPU

- Google : *Un TPU est un circuit intégré propre à une application (ASIC) spécialement développé par Google pour les réseaux de neurones.*

- Intimement lié à l'écosystème Google
  - Disponible uniquement dans le cloud Google
  - Optimisé pour TensorFlow, pas pour PyTorch
  - Conçu pour le stack logiciel Google

Generative AI stack and options, vu par NVIDIA

- Ceci est la vision de NVIDIA
- D’autres options existent, commerciales ou open source, cloud ou on premise
- À chaque étage de la pile, de nombreuses possibilités, toutes en constante évolution
- Tant de choix à faire !
- Compétences fortes requises



Domain-Specific and Generative AI Application Systems

Model Customization, Evaluation, Safety, and Explainability

Model Architecture and Techniques

Systems Optimization

Systems and Applications
Ecosystem: LangChain, LlamaIndex, Weights & Biases
NVIDIA: AI Workbench, NeMo Guardrails

Services and Microservices
Ecosystem: AWS Bedrock, AzureML, Cohere, Google Vertex AI, OpenAI APIs
NVIDIA: NIM, Avatar Cloud Engine (ACE), BioNeMo, NeMo, Picasso

Models
Ecosystem: BLOOM, Llama, Mistral, MPT, OPT, Phi-2, Getty Images AI Generator, Shutterstock 3D Generator
NVIDIA: BioMegatron, Edify, Nemotron

SDKs and Frameworks
Ecosystem: Colossal-AI, HuggingFace Transformers, PyTorch
NVIDIA: A2X, Megatron-LM, NeMo Framework, Riva, Picasso

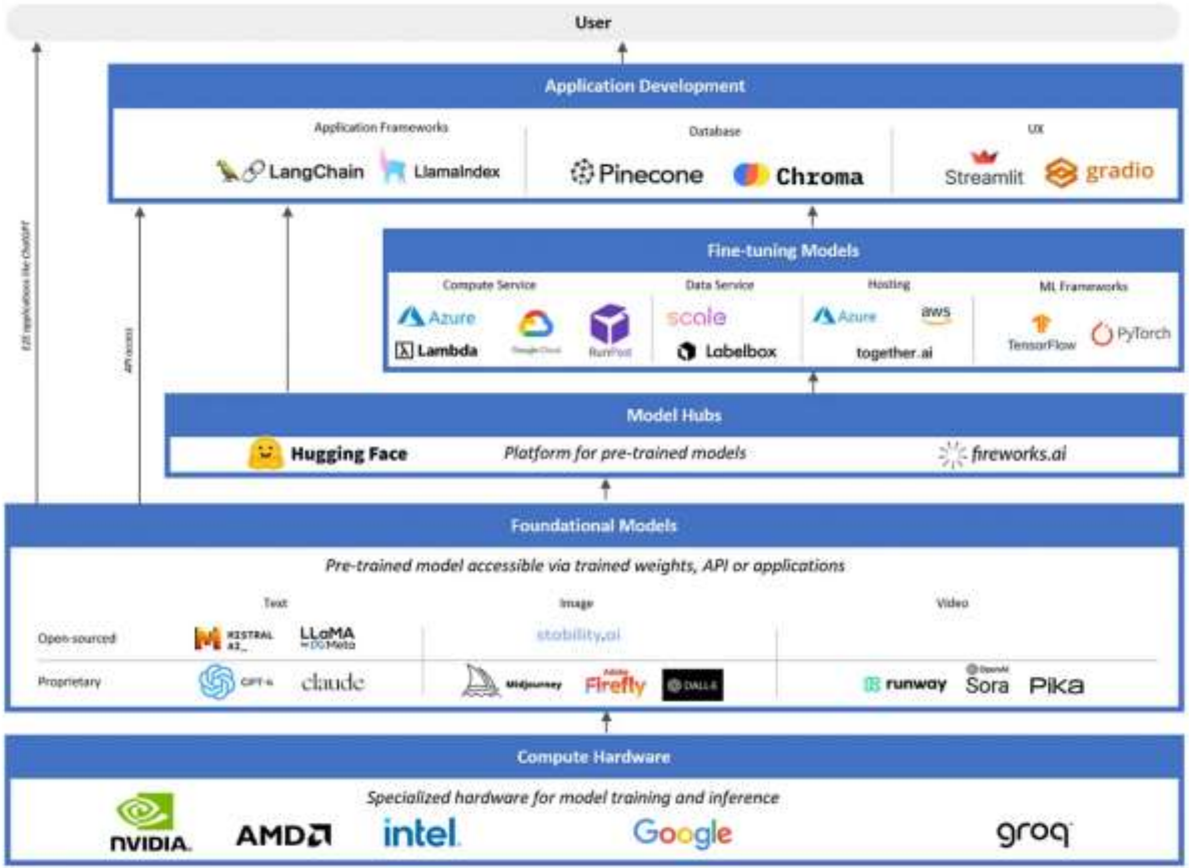
Libraries
Ecosystem: XLA
NVIDIA: CUDA, CUTLASS, CV-CUDA, Megatron-Core, Megatron-LM, NCCL, RAFT, Transformer Engine, TensorRT-LLM, Ray

Management and Orchestration
Ecosystem: Kubernetes, Nephele, Slurm, VMware
NVIDIA: Base Command Platform

Accelerated Infrastructure
NVIDIA: DPUs, GPUs, InfiniBand

https://developer.nvidia.com/topics/ai/generative-ai?

Open source Generative AI stack



User

Application Development

Fine-tuning Models

Model Hubs

Platform for pre-trained models

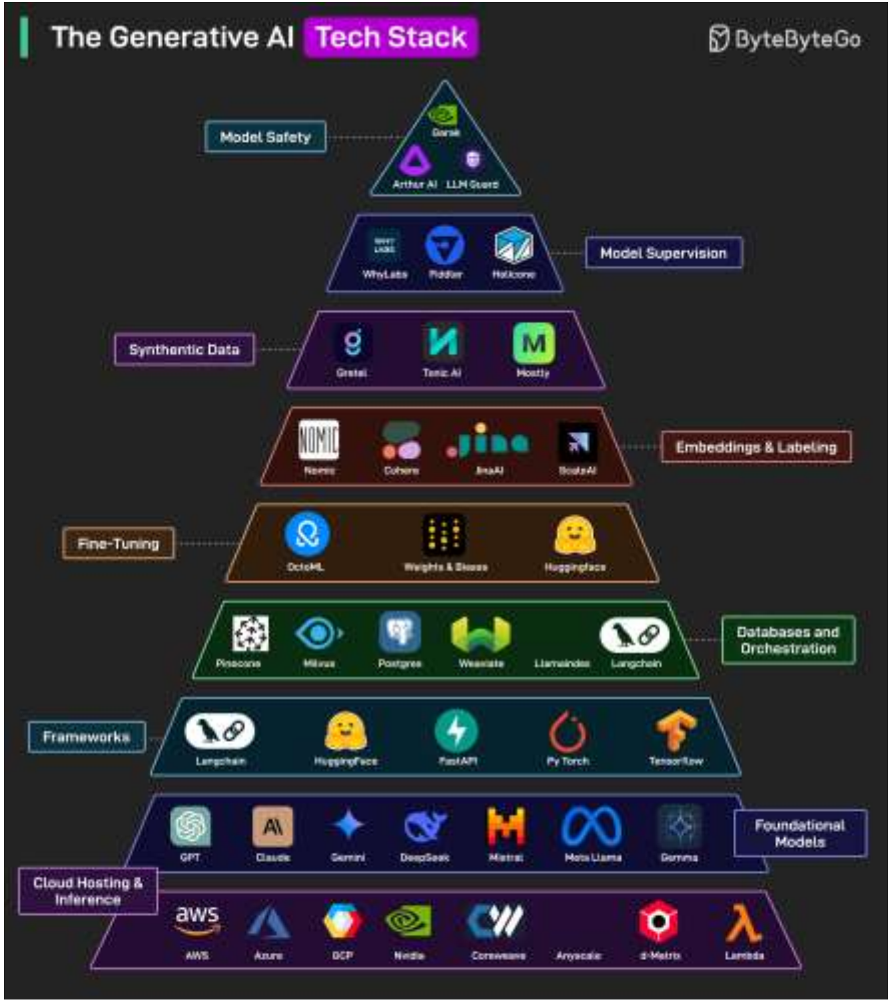
Foundational Models

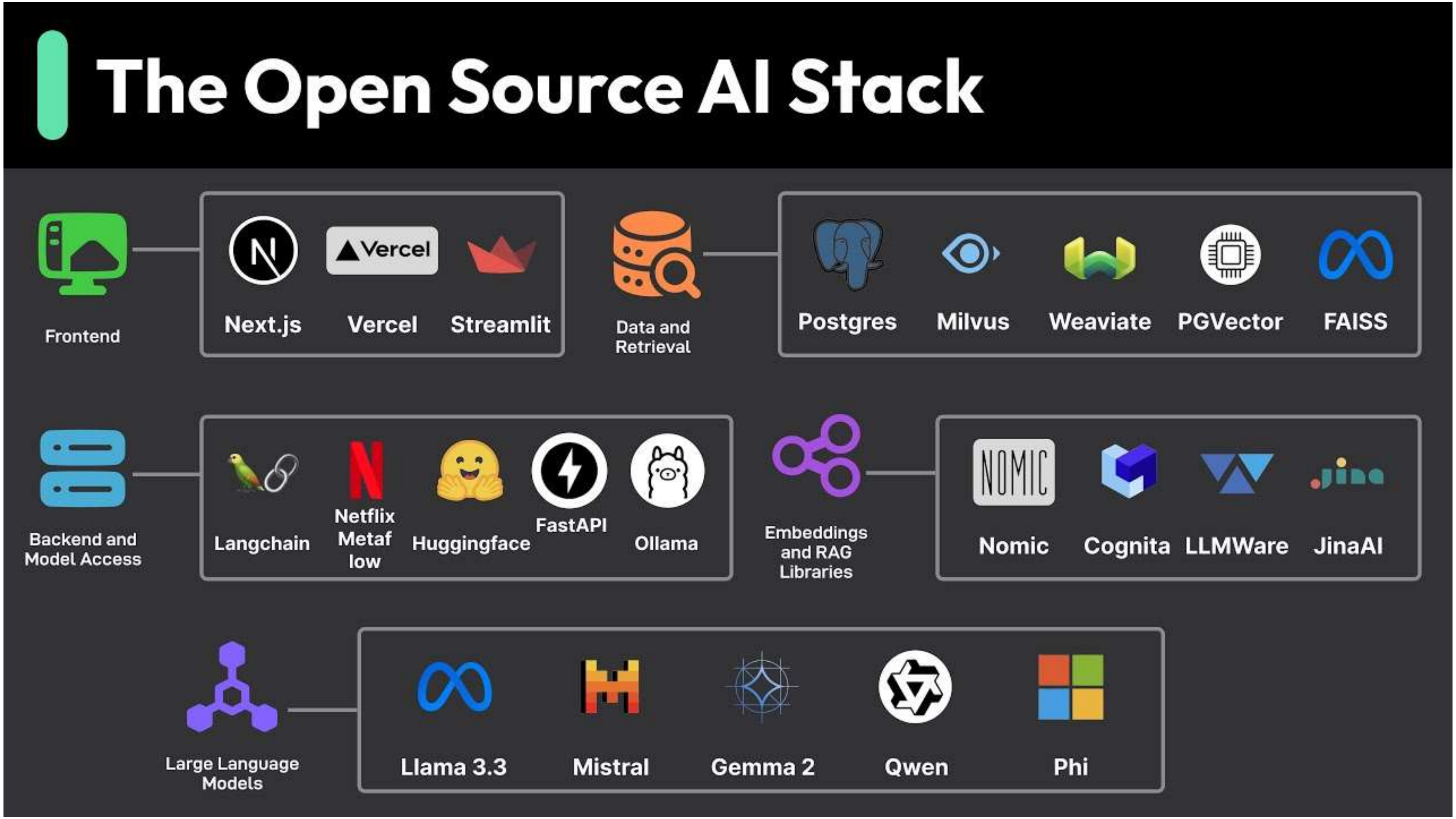
Pre-trained model accessible via trained weights, API or applications

Compute Hardware

Specialized hardware for model training and inference

Open source Generative AI stack





The Open Source AI Stack

Frontend — Next.js, Vercel, Streamlit

Data and Retrieval — Postgres, Milvus, Weaviate, PGVector, FAISS

Backend and Model Access — Langchain, Netflix Metaf low, Huggingface, FastAPI, Ollama

Embeddings and RAG Libraries — Nomic, Cognita, LLMWare, JinaAI

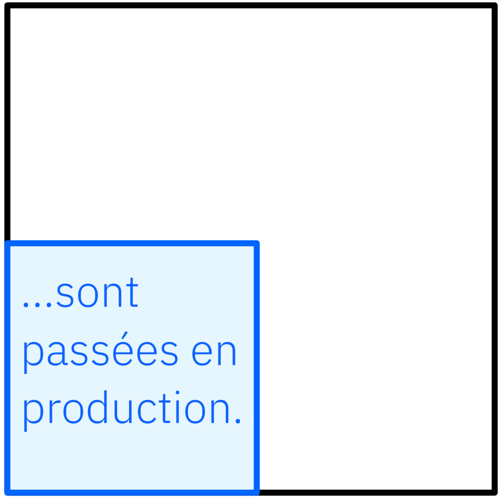
Large Language Models — Llama 3.3, Mistral, Gemma 2, Qwen, Phi

Le résultat ?

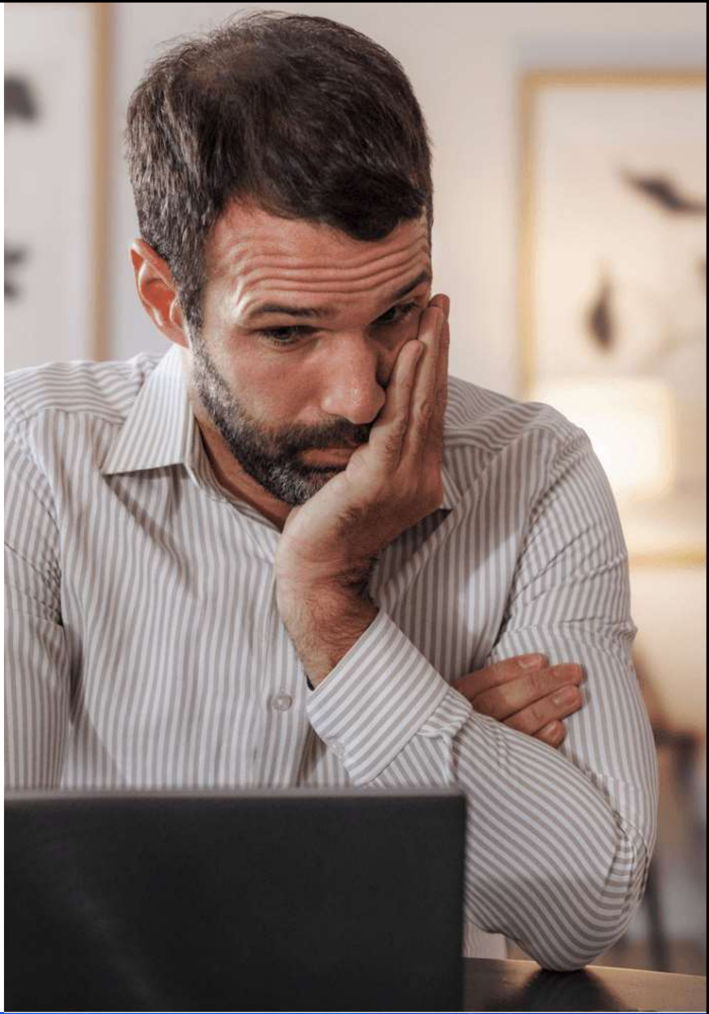
98% des enterprises ont expérimenté l’IA générative[1].

Mais seulement…

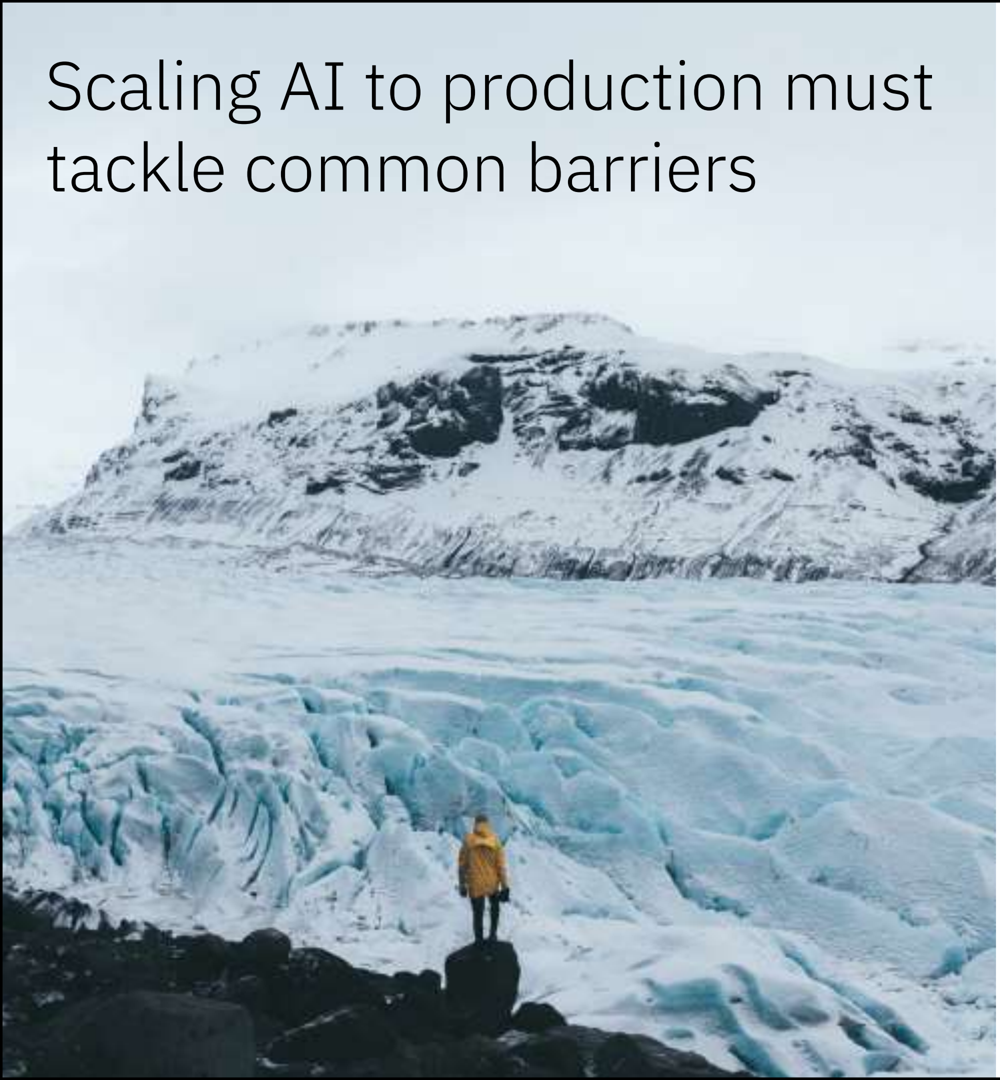
26%[1]



…sont passées en production.



Scaling AI to production must tackle common barriers



3 TOP BARRIERS[1]

- Confidentialité et sécurité des données et réglementations (57%)
- Compétences (35%)
- Complexité (22%)

(1) These insights are drawn from the IBM Institute of Business Value’s “Global AI Adoption Index 2023,” published in January 2024 and Deloitte, “The State of Generative AI in the Enterprise”, 2024.

Confidentialité des données

- **Barrière** : Confidentialité des données (57%) et confiance / transparence (43%) sont les les plus fort inhibiteurs pour l'IA générative[1]
- **Contexte** : Protéger les informations d'identification personnelles (56.6%) et respecter les réglementations (46.0%)[2]
- **Impact** : 92% des projets IA tournent là où sont les données – seulement 16% dans les clouds publics[1]

Compétences

Barrière : le manque de compétences pour implémenter l’IA generative Bloque 35% des initiatives[1]

Contexte : les besoins en compétences

- Sécurité des données (25%),
- technologie (24%),
- littératie des données (20%),
- data science/statistiques (20%)[2]

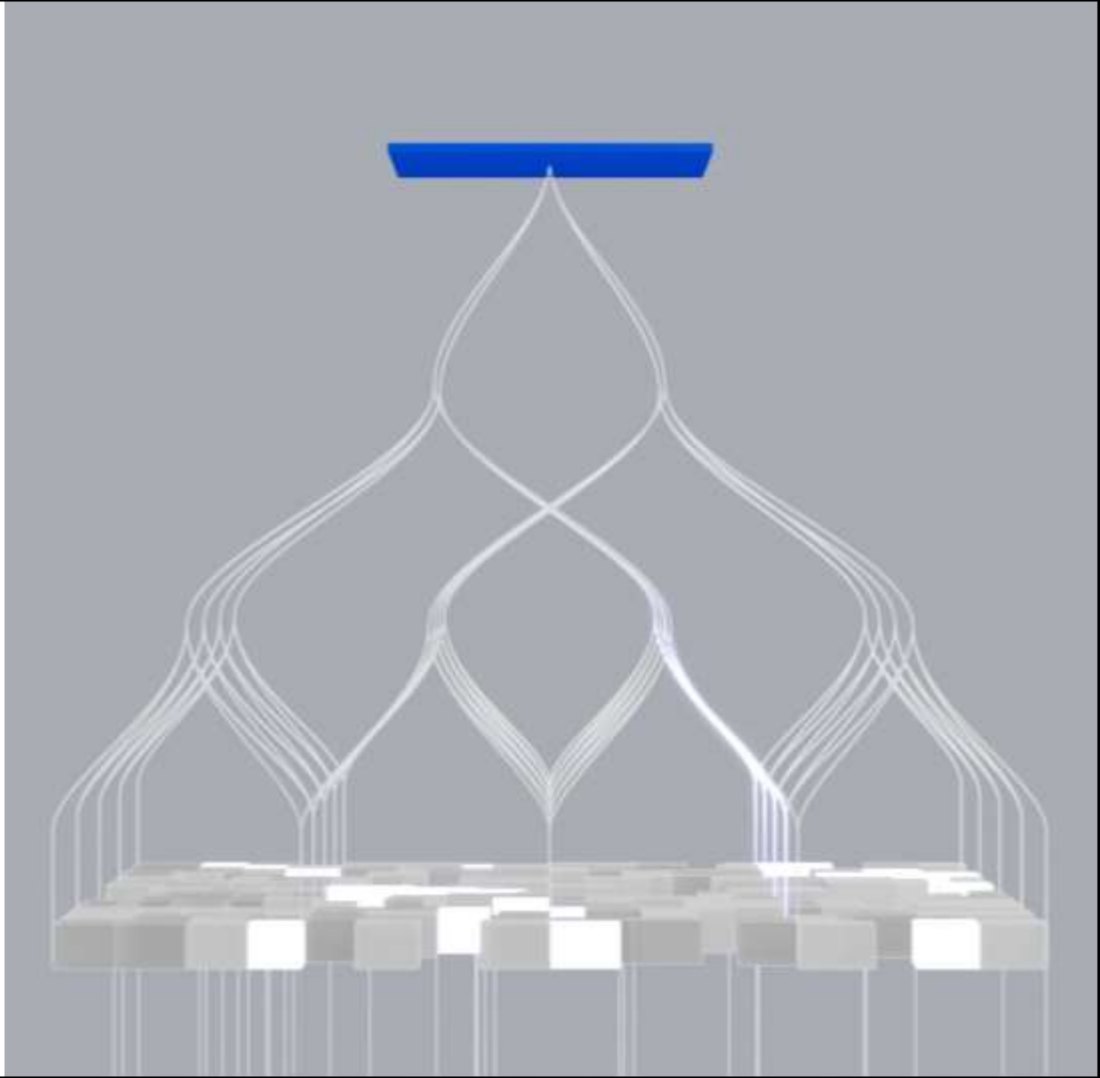
Impact: Les organisations tentent d’acheter (30%), outsourcer (28%), construire (21%), ou panacher ces options (19%)[3]

Complexité

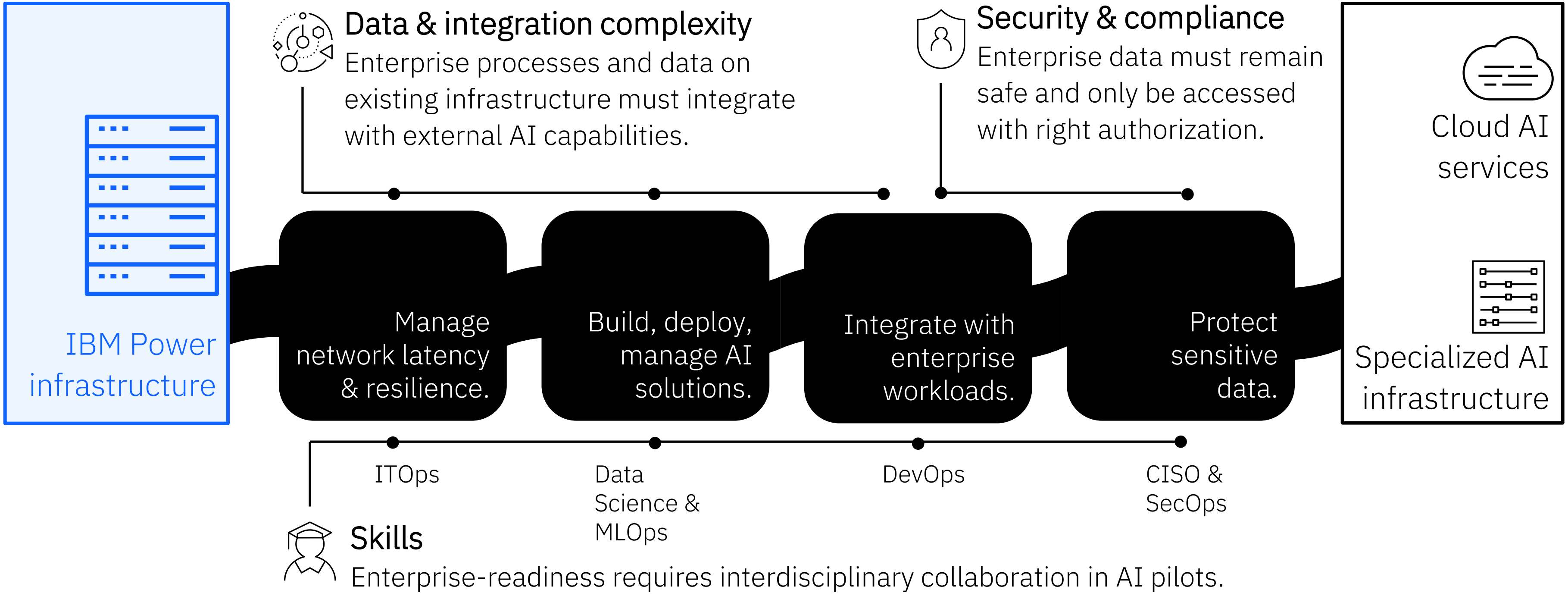
- **Barrières** : 22% des organisations disent que les projets d’IA sont trop complexes ou trop difficiles à intégrer et à faire passer à l’échelle[1]

- **Contexte** : le passage à l’échelle demande :
  - Une integration sans faille avec l’infrastructure et l’architecture de données existante
  - Une infrastructure scalable et capable de traitement en temps réel[2]

- **Impact** : les organisations tentent des partenariats (47%), des connexions entre l’IT et les équipes métiers (43%), et des montées en competence de développement (34%)[2]



Core enterprise workflows get the largest ROI from AI, if obstacles are addressed.



Data & integration complexity

Enterprise processes and data on existing infrastructure must integrate with external AI capabilities.

Security & compliance

Enterprise data must remain safe and only be accessed with right authorization.

IBM Power infrastructure

Manage network latency & resilience.

Build, deploy, manage AI solutions.

Integrate with enterprise workloads.

Protect sensitive data.

Cloud AI services

Specialized AI infrastructure

ITOps

Data Science & MLOps

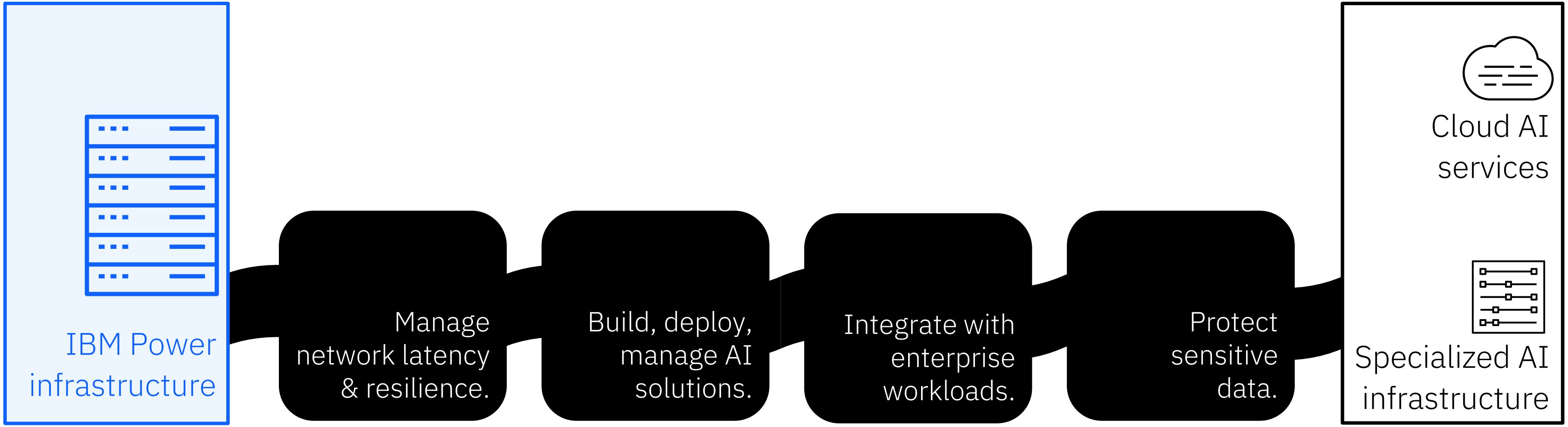
DevOps

CISO & SecOps

Skills

Enterprise-readiness requires interdisciplinary collaboration in AI pilots.

…so, what if all this now comes out-of-the-box?



IBM Power infrastructure

Manage network latency & resilience.

Build, deploy, manage AI solutions.

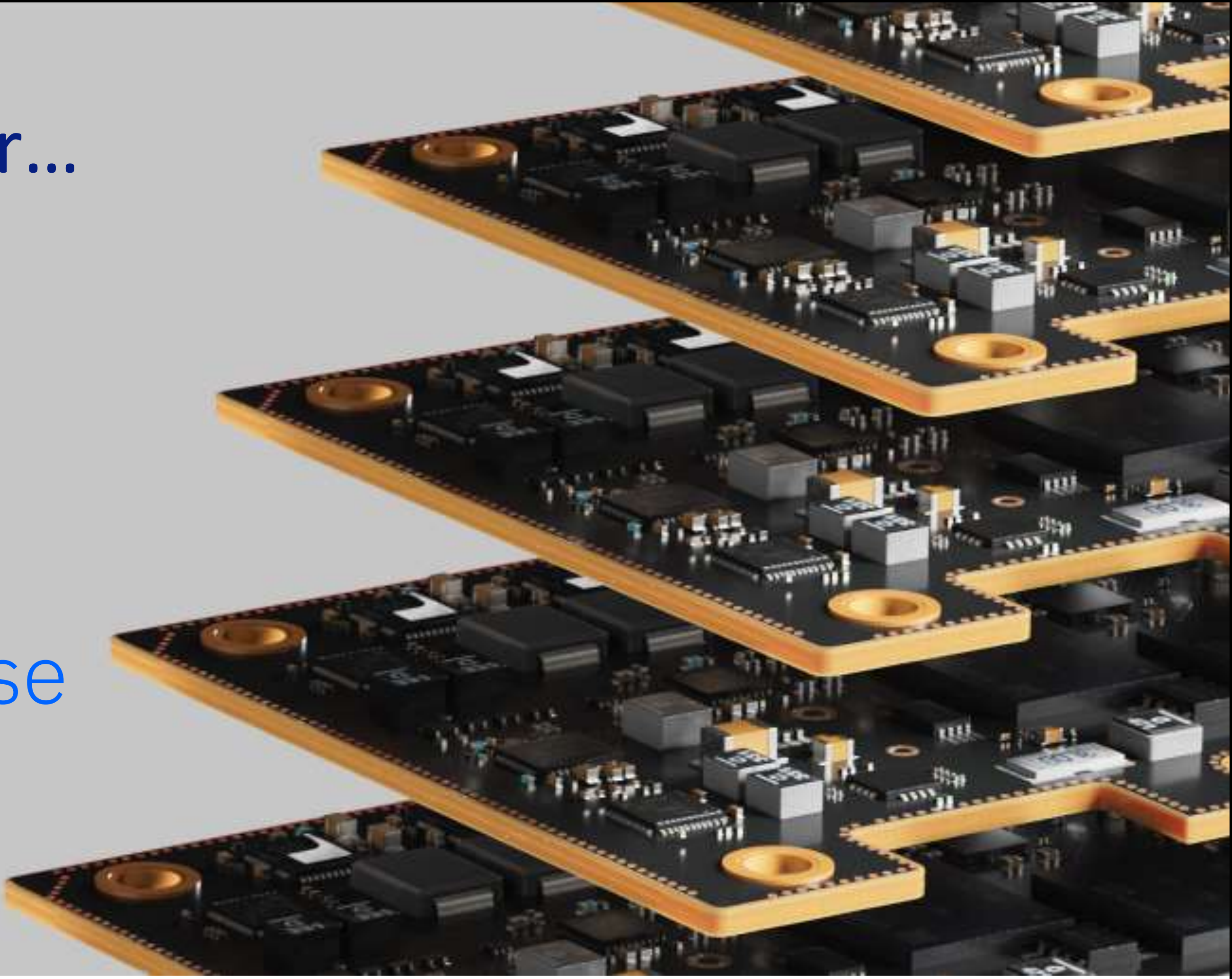
Integrate with enterprise workloads.

Protect sensitive data.

Cloud AI services

Specialized AI infrastructure

Introducing IBM Spyre™ for Power…



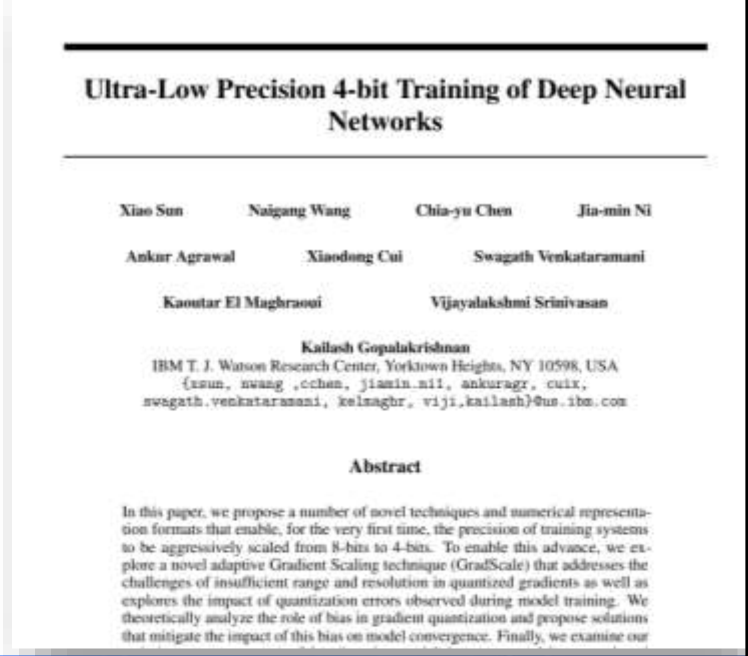
IA clé en main
pour l’enterprise

Size matters… les formats des nombres

Format	Taille (bits)	Signe	Exposant (bits)	Mantisse (bits)	Plage typique	Usage principal
FP64	64	1	11	52	de 2.23×10^{-308} à environ 1.8×10^{308}	**calculs de très haute précision (scientifique, financier, technique)**
FP32	32	1	8	23	1.18×10−38 à environ 3.4×10^{38}	**Calcul scientifique, IA précise**
FP16	16	1	5	10	±65 504	**IA, images, Quantification**
FP8	8	1	4 ou 5	2 ou 3	E4M3 : de 3.9×10^{-3} à 240 E5M2 : de 1.5×10^{-5} à 65 504	**IA, Quantification, Edge**
INT8	8	1	N/A	N/A	-128 à 127	**Quantification, calcul entier**
INT16	16	1	N/A	N/A	-32 768 à 32 767	**Entiers, indices**
INT32	32	1	N/A	N/A	-2 147 483 648 à 2 147 483 647	**Entiers, calculs généraux**

Quantification des modèles

- Inférence plus rapide
- Efficacité matérielle
- Augmentation de l’efficacité énergétique
- Compatibilité élargie



Ultra-Low Precision 4-bit Training of Deep Neural Networks

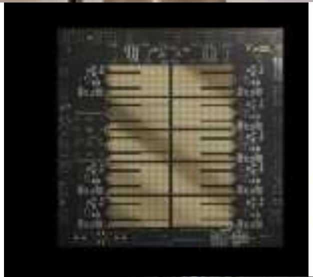
Xiao Sun Naigang Wang Chia-yu Chen Jia-min Ni
Ankur Agrawal Xiaodong Cui Swagath Venkataramani
Kaoutar El Maghraoui Vijayalakshmi Srinivasan
Kailash Gopalakrishnan
IBM T. J. Watson Research Center, Yorktown Heights, NY 10598, USA
{xsun, nwang ,cchen, jiamin.ni1, ankuragr, cuix,
swagath.venkataramani, kelmaghr, viji,kailash}@us.ibm.com

Abstract

In this paper, we propose a number of novel techniques and numerical representation formats that enable, for the very first time, the precision of training systems to be aggressively scaled from 8-bits to 4-bits. To enable this advance, we explore a novel adaptive Gradient Scaling technique (GradScale) that addresses the challenges of insufficient range and resolution in quantized gradients as well as explores the impact of quantization errors observed during model training. We theoretically analyze the role of bias in gradient quantization and propose solutions that mitigate the impact of this bias on model convergence. Finally, we examine our

IBM Albany AI Hardware Center

- launched in February 2019, with large initial investments from IBM, SUNY Polytechnic Institute, and the state of New York.
- Founding members included Samsung, Synopsys, Applied Materials, and Tokyo Electron Limited (TEL).
- University at Albany are also crucial partners, and a sizable portion of the AI Hardware Center's work happens at the Albany NanoTech Complex.
- The IBM Center’s goal is to develop the next generation of chips, systems, and software to support the future of AI.



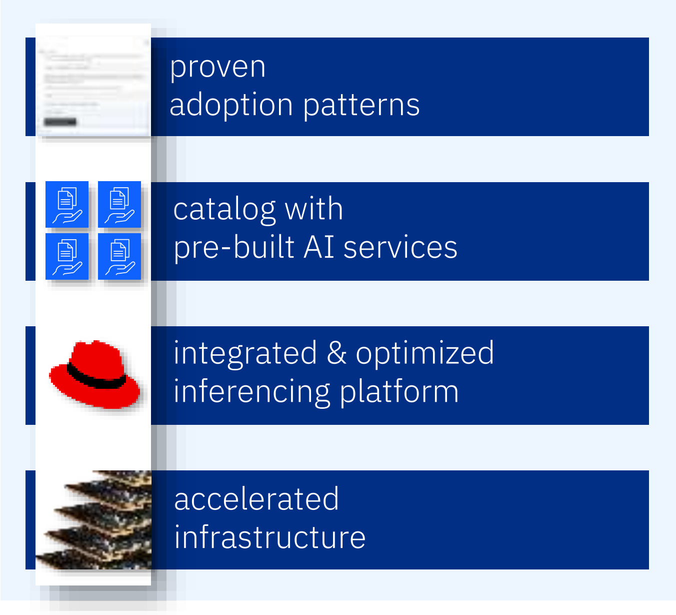
IBM Spyre™ Accelerator - AIU

- IBM AIU : Artificial Intelligence Unit
- ASIC dédié à l'inférence
- 300+ TOPS
- 75W
- PCIe gen5 x16 adapter
- 32 cœurs gravure 5 nm
- 128GB de mémoire Low Power DDR5
- Cluster de 8 cartes
- Vue unifiée par le firmware du cluster : 1To de RAM et 1,6To/s de bande passante mémoire

GPU Model	CUDA Cores	Tensor Cores	Power Usage (W)	First Availability	Memory Size	Tensor FP16 (TFLOPS)
NVIDIA L4	7,424	240	**72**	**Mar 2023**	**24 GB** DDR6	121*

IBM Spyre™ for Power

Turnkey AI for enterprise workloads.



proven adoption patterns

catalog with pre-built AI services

integrated & optimized inferencing platform

accelerated infrastructure

1 click ...to install AI services from the IBM-supported catalog.[1]

1 configuration ...to move AI services of the IBM-supported catalog between IBM Power & IBM Power Virtual Server.[2]

>8 million/hour ...document embeddings for knowledge base integration using IBM Spyre™ Accelerator for Power with batch and prompt sizes of 128.[3]

Disclaimer: 1: AI services of the IBM-supported catalog are delivered as one or a set of containers that can be deployed with a single deployment command. The provided UI for the catalog executes such commands in the backend based on a single click within the UI page of the respective AI service. 2: A single configuration is enabled by exposed industry standard APIs to decouple services at the top and the backing inferencing service for all AI services that are part of the IBM-supported catalog. Any service that requires AI inferencing capabilities can connect inferencing services that provide OpenAI API or watsonx.ai API compliant inferencing endpoints (Spyre endpoint, RH AI Inferencing Server, IBM Cloud, OpenAI, Azure, AWS, GCP, ...). Services can run either on IBM Power or on IBM Power Virtual Server. 3: Based upon internal testing running 1M unit data set with prompt size 128, batch size 128 using 1-card container. Individual results may vary based on workload size, use of storage subsystems and other conditions.

Spyre Security

Control: 100% control to unlock trusted AI services where enterprise data resides (on-premises or IBM PowerVS).

Air-gapped: Install & operate AI services without internet access.

Support: One-stop-shop supported & integrated enterprise AI stack.

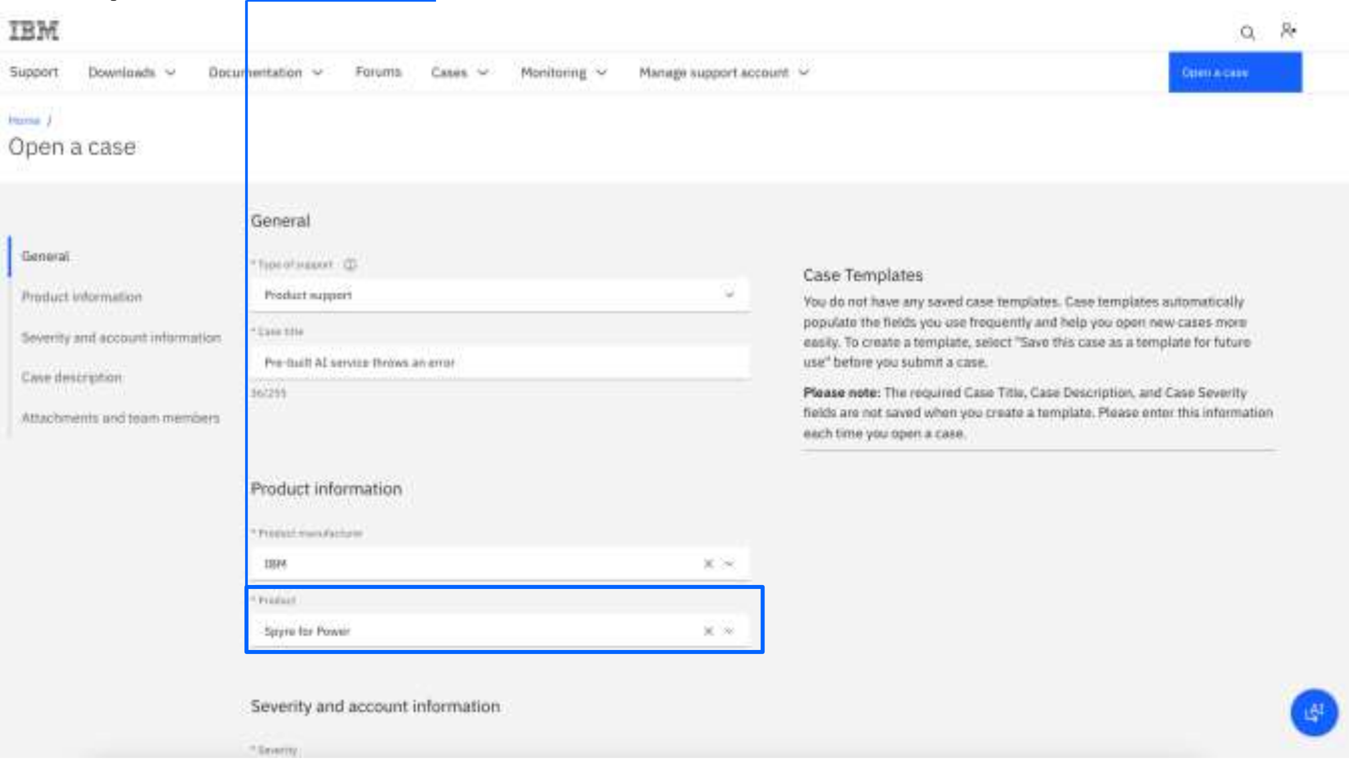
Network security: 100% native IBM Power ecosystem without data exposure to any network between Spyre drawer and IBM Power server.

Enterprise Support

ibm.com/support

- File tickets for any component of the Spyre stack to get help!

Example 1: Software issue



Open a case

General
Product information
Severity and account information
Case description
Attachments and team members

General

* Type of support

Product support

* Case title

Pre-built AI service throws an error

36/255

Product information

* Product manufacturer

IBM

* Product

Spyre for Power

Severity and account information

* Severity

Case Templates

You do not have any saved case templates. Case templates automatically populate the fields you use frequently and help you open new cases more easily. To create a template, select "Save this case as a template for future use" before you submit a case.

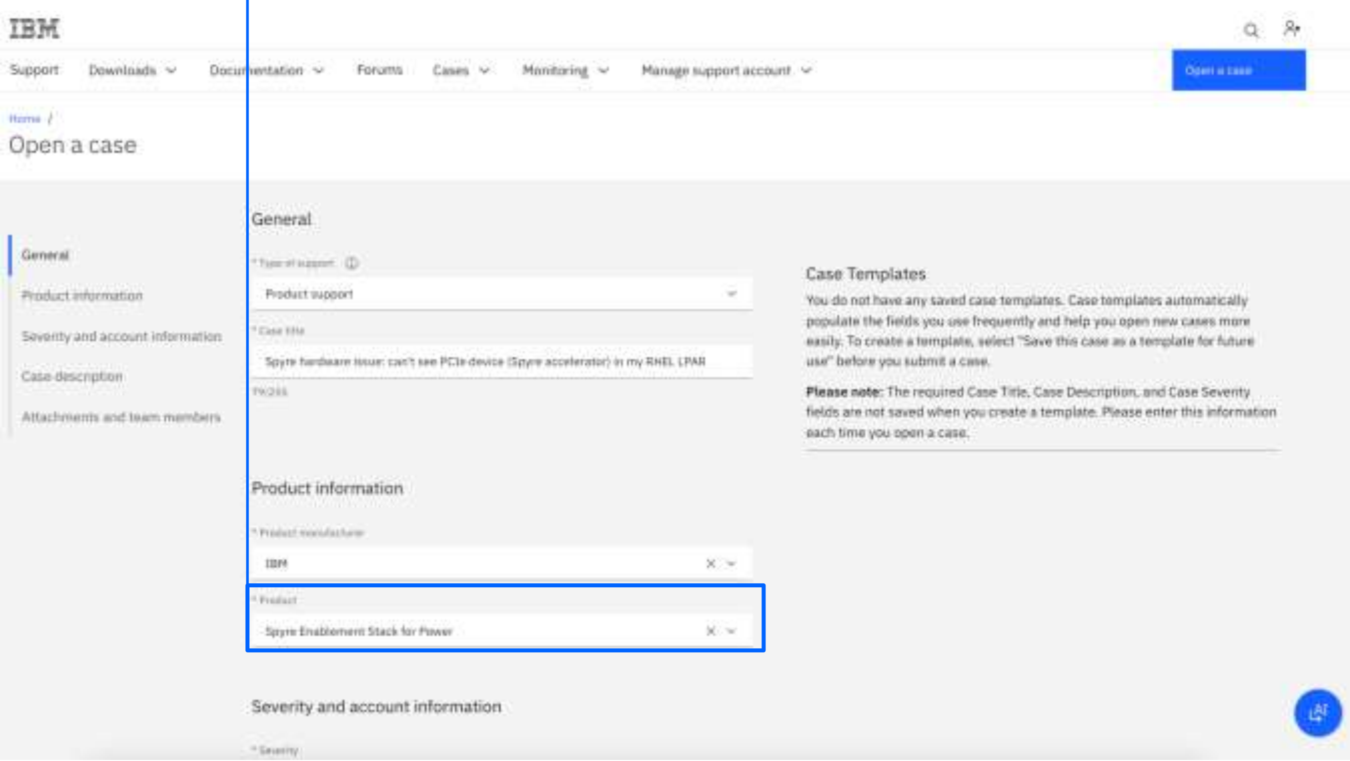
Please note: The required Case Title, Case Description, and Case Severity fields are not saved when you create a template. Please enter this information each time you open a case.

Enterprise Support

ibm.com/support

- File tickets for **any** component of the Spyre stack to get help!

Example 2: Hardware issue



Open a case

General

* Type of support: Product support

* Case title: Spyre hardware issue: can't see PCIe device (Spyre accelerator) in my RHEL LPAR

79/255

Product information

* Product manufacturer: IBM

* Product: Spyre Enablement Stack for Power

Severity and account information

* Severity

Case Templates

You do not have any saved case templates. Case templates automatically populate the fields you use frequently and help you open new cases more easily. To create a template, select "Save this case as a template for future use" before you submit a case.

Please note: The required Case Title, Case Description, and Case Severity fields are not saved when you create a template. Please enter this information each time you open a case.

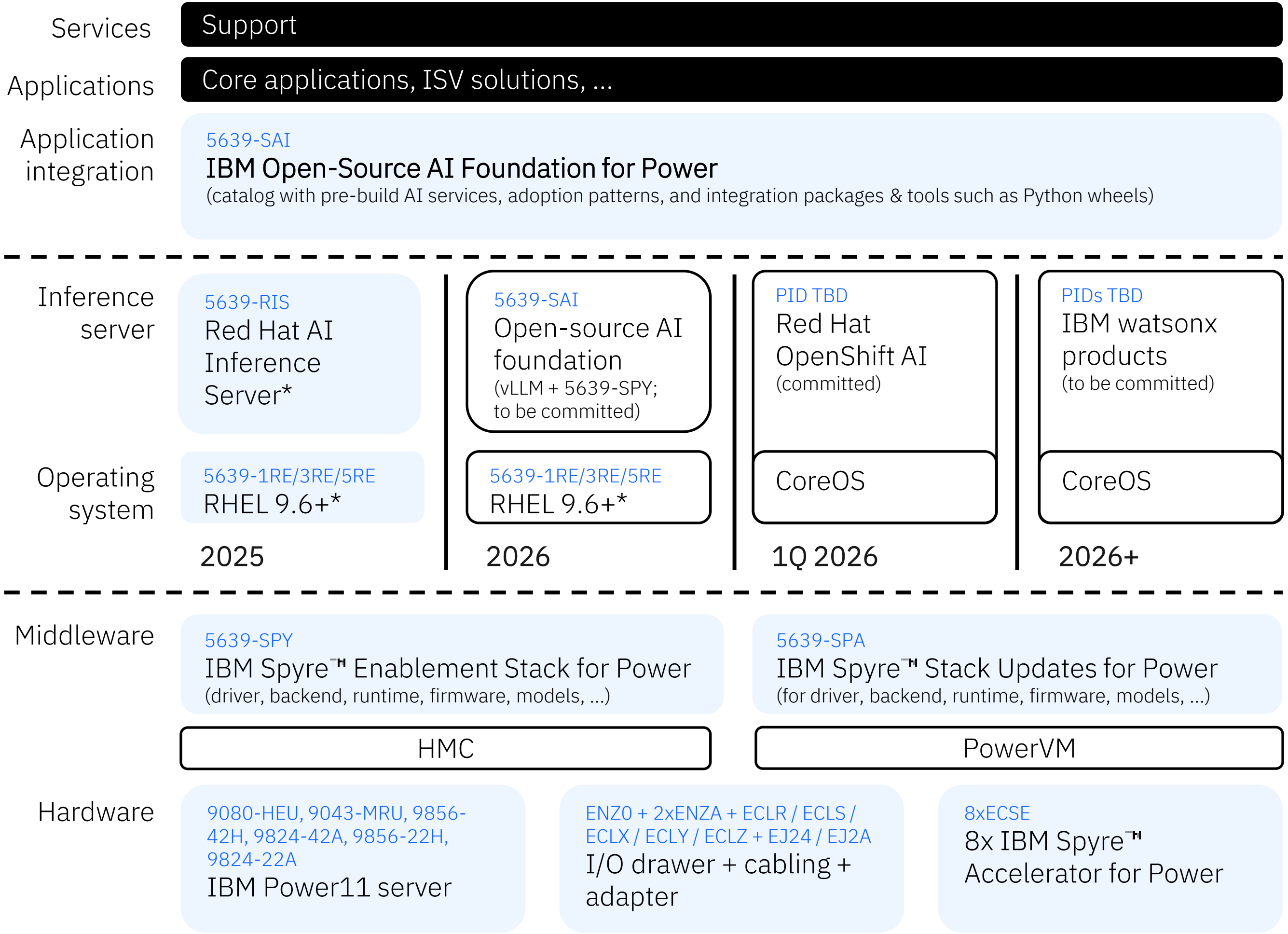
Spyre Stack

external

mandatory

optional

*removal allowed if purchased via Red Hat. Support for those components will then come via Red Hat.



Services

Support

Applications

Core applications, ISV solutions, ...

Application integration

5639-SAI
IBM Open-Source AI Foundation for Power
(catalog with pre-build AI services, adoption patterns, and integration packages & tools such as Python wheels)

Inference server

Operating system

2025

5639-RIS
Red Hat AI Inference Server*

5639-1RE/3RE/5RE
RHEL 9.6+*

2026

5639-SAI
Open-source AI foundation
(vLLM + 5639-SPY; to be committed)

5639-1RE/3RE/5RE
RHEL 9.6+*

1Q 2026

PID TBD
Red Hat OpenShift AI
(committed)

CoreOS

2026+

PIDs TBD
IBM watsonx products
(to be committed)

CoreOS

Middleware

5639-SPY
IBM Spyre™ Enablement Stack for Power
(driver, backend, runtime, firmware, models, ...)

5639-SPA
IBM Spyre™ Stack Updates for Power
(for driver, backend, runtime, firmware, models, ...)

HMC

PowerVM

Hardware

9080-HEU, 9043-MRU, 9856-42H, 9824-42A, 9856-22H, 9824-22A
IBM Power11 server

ENZ0 + 2xENZA + ECLR / ECLS / ECLX / ECLY / ECLZ + EJ24 / EJ2A
I/O drawer + cabling + adapter

8xECSE
8x IBM Spyre™ Accelerator for Power

IBM Spyre™ Solutions catalog

Optimized

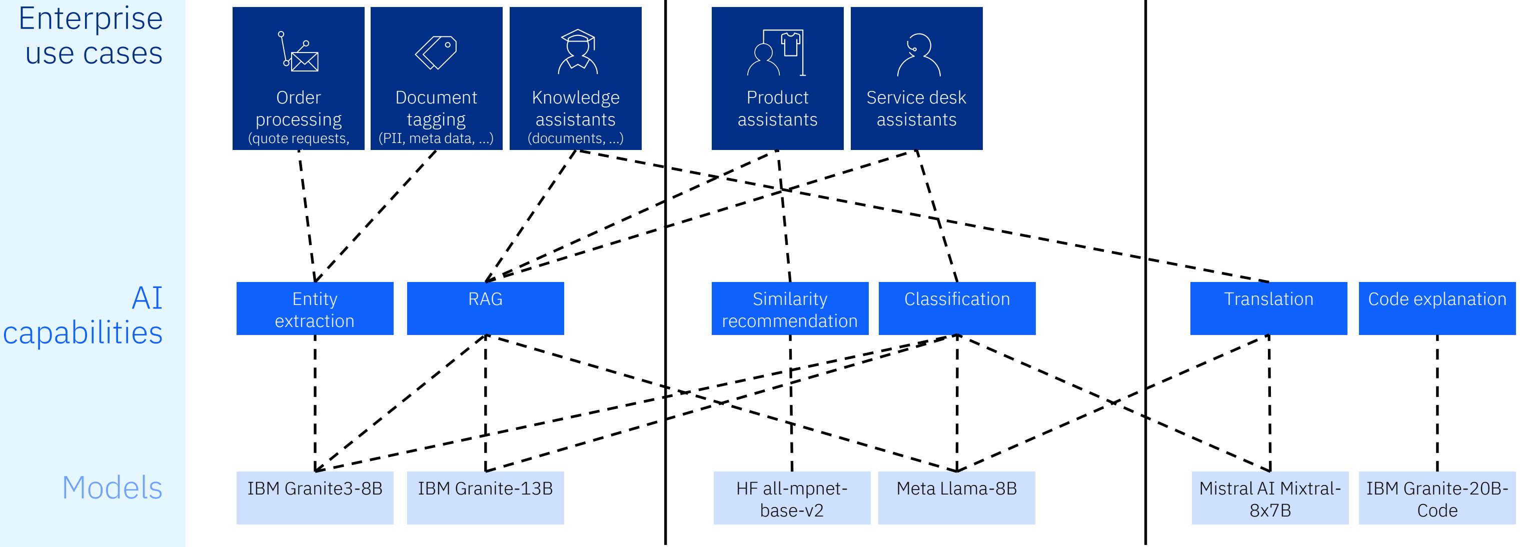
IBM has optimized accuracy & performance for enterprise use cases.

Tested

Functionally working and exploring for enterprise use cases.

Available

DIY exploration and experimentation of AI capabilities.



Final determination of use cases, capabilities, and models subject to change until GA.

Initial use cases with IBM Spyre™ for Power

(currently validated in Tech Preview; more to come)

Technical AI capabilities

- ▢ Traditional AI
- 🟦 Generative AI
- ■ Agents and assistants

Cross-industry

ITOps | Development

- ■ Service desk assistant
- ■ Code assistants (RPG, Ansible, ...)
- ■ Maintenance & compliance agent
- 🟦 Generate IT reports
- ▢ Forecast & plan capacity
- ▢ Predict IT issues
- ▢ Detect & fix incidents

Data & content management

- ■ Knowledge assistant (RAG, ...)
- 🟦 Translation & summarization
- 🟦 Document tagging (PII, meta data, ...)
- 🟦 Document digitalization (manual, invoice,...)
- 🟦 Meeting transcription

Enterprise Resource Planning systems

- ■ BI & HR assistants
- ■ Product assistant
- 🟦 Order processing (quote requests, ...)
- 🟦 Compliance checking (invoices, ...)
- 🟦 Sales insights
- ▢ Predict customer churn
- ▢ Optimize stocks & demands
- ▢ Business intelligence
- ▢ Supply chain forecasting
- ▢ Visual quality inspection

Industry-specific

Finance & banking

- ■ Analyst assistant (frauds, NPA, ...)
- ▢ Anti-money laundering
- ▢ NPA prediction
- ▢ Fraud detection
- ▢ Know Your Customer

Health care

- ■ Medical assistant
- ■ Claims & EHR matching agent
- ▢ Medical transcription
- ▢ Medical image analysis

Insurance

- ■ Claim & policy assistant
- ■ Claims management agent
- 🟦 Customer call summarization
- ▢ Claim fraud detection
- ▢ Underwriting & risk prediction

...and more

Client examples & demos

aligned with core workloads; not restricted to Spyre™.

Cross-industry

ITOps | Development

- System house DACH – IT service desk assistant
- Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

Enterprise Resource Planning

- BI & HR assistant
- Large retailer US – Supply chain forecasting
- Order processing assistant
- Large retailer US – Product sales assistant

Industry-

Banking & Finance

- Demo TechXchange 2024 – Analyst assistant (frauds, NPAs, ...)
- Finacle – Predict NPAs
- Open account agent
- Anti-money laundering

Healthcare

- SHIBUYA – Medical assistant
- Medical transcription assistant
- Claims & EHR matching agent
- Medical image analysis assistant

Insurance

- Demo TechXchange 2025 – Claims & policy management agent
- Predict risk & underwrite assistant

Public

- Gov. client DACH – Private documents assistant
- 360-degree view assistant

...and more

- SEMICON INDIA – Agriculture assistant
- Real estate assistant

...

proven Adoption patterns

- Digital assistant (RAG, ...)
- Data & content management
- Recommender system
- Deep process integration
- Fraud detection
- Forecasting
- Image & video analytics

...

pre-built AI Services

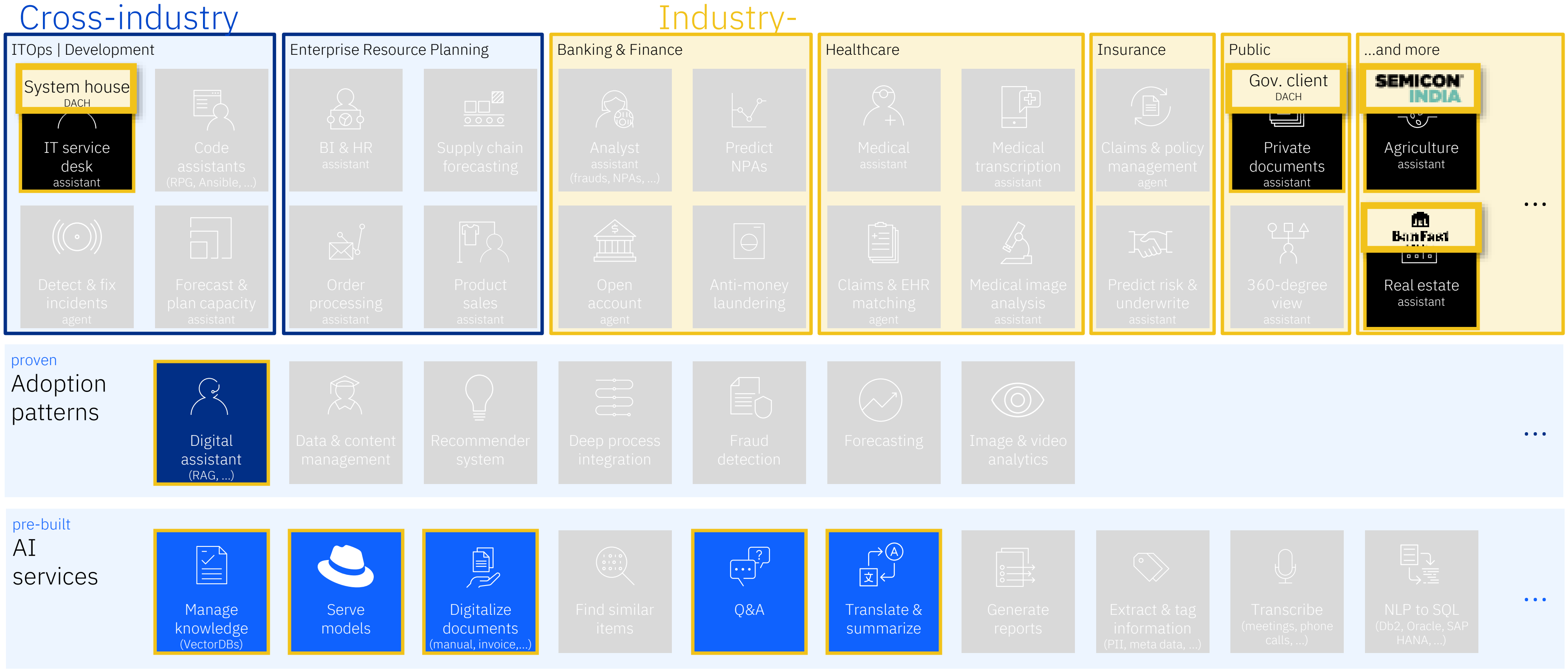
Open Source containers available from in IBM cloud registry

- Manage knowledge (VectorDBs)
- Serve models
- Digitalize documents (manual, invoice,...)
- Find similar items
- Q&A
- Translate & summarize
- Generate reports
- Extract & tag information (PII, meta data, ...)
- Transcribe (meetings, phone calls, ...)
- NLP to SQL (Db2, Oracle, SAP HANA, ...)

...

GA 1 scope

additional services & patterns to follow post GA 1 regularly.



Cross-industry

ITOps | Development

- System house DACH: IT service desk assistant
- Code assistants (RPG, Ansible, ...)
- Detect & fix incidents agent
- Forecast & plan capacity assistant

Enterprise Resource Planning

- BI & HR assistant
- Supply chain forecasting
- Order processing assistant
- Product sales assistant

Industry-

Banking & Finance

- Analyst assistant (frauds, NPAs, ...)
- Predict NPAs
- Open account agent
- Anti-money laundering

Healthcare

- Medical assistant
- Medical transcription assistant
- Claims & EHR matching agent
- Medical image analysis assistant

Insurance

- Claims & policy management agent
- Predict risk & underwrite assistant

Public

- Gov. client DACH: Private documents assistant
- 360-degree view assistant

...and more

- SEMICON INDIA: Agriculture assistant
- Real estate assistant
- ...

proven Adoption patterns

- Digital assistant (RAG, ...)
- Data & content management
- Recommender system
- Deep process integration
- Fraud detection
- Forecasting
- Image & video analytics
- ...

pre-built AI services

- Manage knowledge (VectorDBs)
- Serve models
- Digitalize documents (manual, invoice,...)
- Find similar items
- Q&A
- Translate & summarize
- Generate reports
- Extract & tag information (PII, meta data, ...)
- Transcribe (meetings, phone calls, ...)
- NLP to SQL (Db2, Oracle, SAP HANA, ...)
- ...

Sizing
Digital assistants

AI services:
- Digitize documents
- Knowledge management
- Q&A
- Serve models

Backend services:
- Model server
  (Red Hat AI Inference Server)
- VectorDB
  (Milvus)

Starter

Basic assistant for 5 concurrent users.

1 RHEL 9.6+ LPAR with all service requires…

- **Spyre cards**: 8
- **CPU**[1]:
  - Scale-out: 12 cores (via 24 core DCM)
  - Mid-range: 12 cores (via 24 core DCM)
  - High-end: 12 cores (via 12 core SCM)
- **Memory**[2]: 512 GB
- **Storage**[3]: 600 GB

Note: Using "digitize documents" should happen *offline*; before using the "Q&A" service.

Production

Higher accuracy through filtering & reranking.

1 RHEL 9.6+ LPAR with all service requires…

- **Spyre cards**: 8
- **CPU**[1]:
  - Scale-out: 15 cores (via 30 core DCM)
  - Mid-range: 15 cores (via 30 core DCM)
  - High-end: 16 cores (via 16 core SCM)
- **Memory**[2]: 512 GB
- **Storage**[3]: 600 GB

Note: Using "digitize documents" should happen *offline*; before using the "Q&A" service.

Scaling

More concurrent users & redundancies.

- **Create additional LPARs** using the starter/production configurations.
- **2x the resources will increase concurrent users by 2x.**
- **Add a dedicated production LPAR for running "Digitize documents"** ***online***, allowing to run digitization tasks in parallel to the Q&A service.

[1] Optimal performance with higher core counts per Power11 chip (high-end & mid-range systems have options for 15 and 12, respectively); NUMA-aligned LPAR with SMT 2; see https://community.ibm.com/community/user/blogs/sebastian-lehrig/2024/03/26/sizing-for-ai.
[2] Use fully populated DIMMs for optimal performance (16x64 GB DIMMs is minimal configuration for optimal performance).
[3] Recommended to use distributed file systems such as NFS storage for VectorDB.

Sizing

Deep process integration

AI services:

- Extract & tag information
- Serve models

Backend services:

- Model server
  (Red Hat AI Inference Server)

Starter

Extraction capabilities to test baseline throughput.

1 RHEL 9.6+ LPAR with all service requires…

- **Spyre cards**: 4
- **CPU**[1]:
  - Scale-out: 6 cores (via 24 core DCM)
  - Mid-range: 6 cores (via 24 core DCM)
  - High-end: 6 cores (via 12 core SCM)
- **Memory**[2]: 256 GB
- **Storage**: 300 GB

Production

2x higher throughput.

1 RHEL 9.6+ LPAR with all service requires…

- **Spyre cards**: 8
- **CPU**[1]:
  - Scale-out: 15 cores (via 30 core DCM)
  - Mid-range: 15 cores (via 30 core DCM)
  - High-end: 16 cores (via 16 core SCM)
- **Memory**[2]: 512 GB
- **Storage**: 600 GB

Scaling

Higher throughput & redundancies.

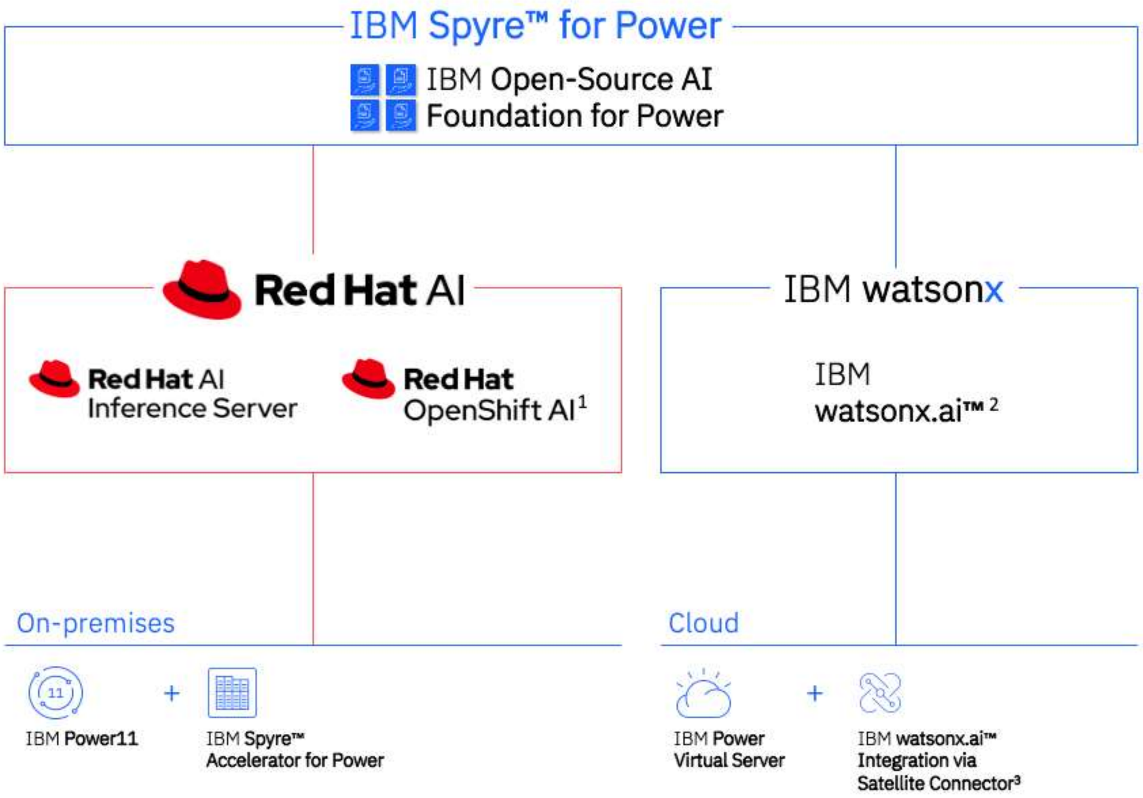
- **Create additional LPARs** using the starter/production configurations.
- **2x the resources will increase throughput by 2x.**

1: LPAR with SMT 2.
2: Use fully populated DIMMs for optimal performance (16x64 GB DIMMs is minimal configuration for optimal performance).

Simplify with turnkey AI…

…integrated with trusted, consistent, and comprehensive inferencing options…

…deployed where data is secure.



IBM Spyre™ for Power

IBM Open-Source AI Foundation for Power

Red Hat AI

Red Hat AI Inference Server

Red Hat OpenShift AI[1]

IBM watsonx

IBM watsonx.ai™ [2]

On-premises

IBM Power11 + IBM Spyre™ Accelerator for Power

Cloud

IBM Power Virtual Server + IBM watsonx.ai™ Integration via Satellite Connector[3]

[1] GA with Spyre 1Q'26

[2] Integration of IBM Spyre™ for Power catalog of AI services targeted for 2026

[3] https://cloud.ibm.com/docs/powervs-watsonx-toolkit?topic=powervs-watsonx-toolkit-powervs-watsonx-ra

Spyre n'est pas un GPU…

Spyre est

- une offre intégrée de matériel et de logiciel,
- hautement efficace,
- pour délivrer en production,
- des solutions d'IA validées et supportées,
- clef en main,
- en un clic,
- au plus près des données, en sécurité.

Automatiser le déploiement d’un environnement IA sur POWER Avec Project PIM

Comment déployer simplement un environnement IA dans une LPAR VM POWER ?

https://github.com/IBM/project-pim

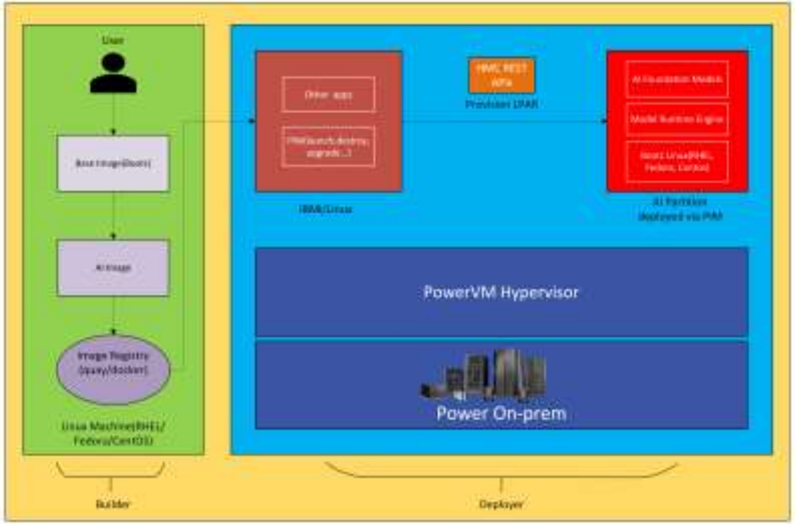
PIM has 2 personas, namely the builder and the deployer :

- **Builder:** Someone who builds a bootable AI container image to bring up the AI stack with the deployer flow.
- **Deployer:** Someone who deploys a PIM solution to bring up the AI stack in IBM core environments.

Key highlights of the PIM solution

- Seamless Update: System updates are automatic if a newer version of the image is publicly available. Otherwise, when the user upgrades via PIM upgrade command with the latest credentials, the system updates are pulled and applied from the configured private registry over a reboot of the system.
- Rollback: bootc preserves the state of the system. In case of a disruption in Updates, the system can be rolled back to a previous version.
- Makes admin's management simple by easing day 2 operations like monitoring, upgrading and managing.
- Provides end-to-end software lifecycle management operations like launch, destroy, update-config, update-compute, rollback and status.
- Provides AI inferencing capability on CPU currently. The intent is to provide inferencing-based accelerators available on the platform as and when they become available.

PIM Architecture



The PIM project enables the spinning up of an AI environment with very little user intervention, adjacent to other workloads running on IBM Power. These workloads might be running on any of the supported operating systems on IBM Power: IBMi, or Linux, as long as they are managed by a Hardware Management Console (HMC).

The PIM solution leverages Bootable Containers (bootc), a modern tool for deploying and configuring immutable Linux systems. PIM provides an end-to-end solution for AI stack installation by creating a Logical Partition (LPAR) with a specified AI stack image. This involves network and storage attachment, and the LPAR is then booted with the configured image.

MERCI